Dirty Comments and Clean Plates: Using customer reviews to predict restaurant’s health inspection results

Claire Boyd, Raul Castellanos, Jack Gibson, Benjamin Leiva
(GitHub: https://github.com/claireboyd/dirty_comments_and_clean_plates)

Abstract

The Philadelphia Department of Public Health conducts yearly safety inspections on food facilities, focusing on preventing foodborne illnesses and educating operators on proper food handling. However, this routine may lead to missed chances to address health and hygiene issues and redundant inspections of compliant establishments. This project has two aims: first, to identify potential violations of health and safety codes by developing a classification model using both inspection data and Yelp reviews and second, to create a model to detect if a restaurant review is fake to help policymakers filter out fake reviews in their modeling approaches.

Introduction

Public health departments are tasked with the critical responsibility of conducting inspections of food facilities to prevent foodborne illnesses and ensure proper food handling practices. However, many health departments face significant resource constraints that limit their ability to inspect all restaurants within their jurisdiction on a regular basis. As a result, departments often resort to randomly selecting a sample of restaurants for inspection each year.

This random inspection approach has several drawbacks. First, it fails to account for factors like inspector travel time and shift schedules, leading to inefficient allocation of limited inspection resources. Additionally, the conditions at a restaurant can change dramatically between the time a complaint is received and when an actual inspection occurs, potentially allowing health code violations to go undetected (Hutton).

On average, health inspectors are only able to complete 3 to 4 inspections per day, with each inspection ranging from one hour to several hours depending on the circumstances (Krishna). Given these constraints, utilizing consumer reviews as an additional data source to identify restaurants potentially in violation of health codes could substantially improve the efficiency and efficacy of the inspection process.

By leveraging crowdsourced consumer feedback, health departments can use review data to flag high-risk establishments for priority inspections. This risk-based inspection model using review data as a filter would complement the existing random inspection system, allowing limited inspection resources to be allocated more strategically (Hutton). The following sections will explore a possible application of consumer reviews into existing health department practices, examining its potential to enhance the efficacy of food safety inspections. Drawing upon insights from existing literature and empirical evidence, we aim to elucidate the benefits and challenges associated with this approach, offering recommendations for its implementation in real-world scenarios.

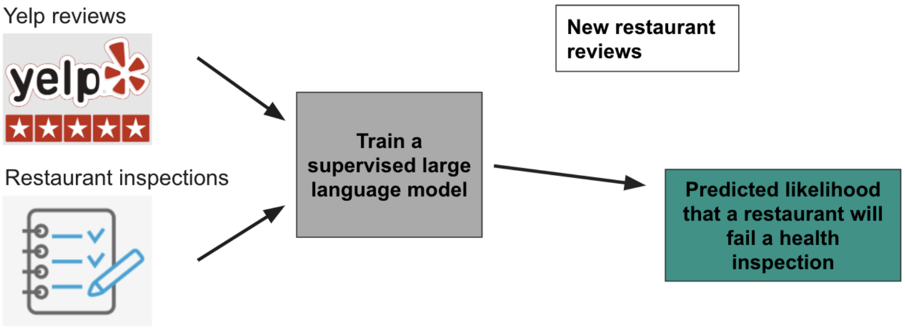
Specifically, we constructed a model to predict whether individual Yelp reviews for restaurants were "fake" or genuinely reflected consumers' experiences. By filtering out fake reviews, we could more accurately gauge the negative sentiment and potential health code issues raised in authentic reviews.

The following sections detail the model development process for detecting fake reviews, explore the efficacy of using consumer review data to augment inspections, and evaluate the potential public health impacts of this approach. Ultimately, by combining AI/ML techniques with crowdsourced data, we aim to enhance the efficiency and effectiveness of regulatory efforts to ensure food safety.

Our task

Our task (**Figure 1)** was to leverage a corpus of Yelp reviews for restaurants that have undergone health inspections in Philadelphia to train a supervised large language model capable of classifying the likelihood that a given restaurant will fail an upcoming health inspection. The inputs to the model were the Yelp review text data and past health inspection results. By training on this data, the model learned to extract relevant signals from the reviews and map them to health code violations identified during inspections. The output was a predicted probability that a restaurant would fail its next health inspection, with this prediction being updated as new reviews for the restaurant became available. This allowed us to prioritize inspections for establishments with the highest predicted risk of health code violations based on consumer feedback.

Figure 1: Diagram of our task


Yelp reviews
yelp
Restaurant inspections
Train a supervised large language model
New restaurant reviews
Predicted likelihood that a restaurant will fail a health inspection

Origin of the data

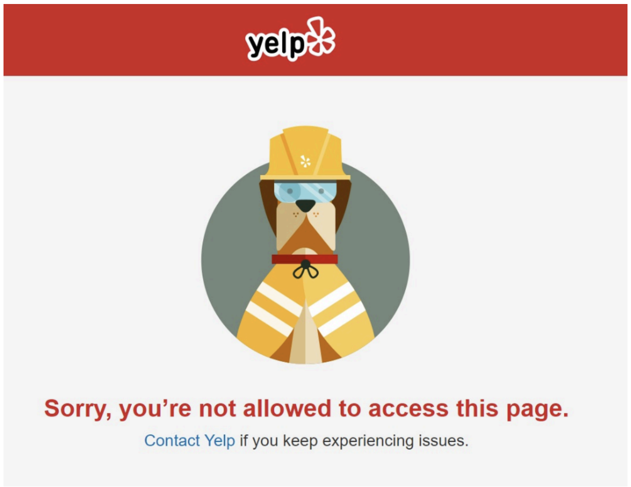
About the Philadelphia inspections

The Environmental Health Services division (EHS) of the Philadelphia Department of Public Health is responsible for conducting routine inspections of food facilities within the city. These inspections typically occur annually and focus on preventing foodborne illness while educating facility operators on proper food handling techniques. Certain establishments, such as hospitals serving vulnerable populations, may undergo more frequent inspections due to their high-risk nature. EHS also conducts reinspections, complaint-based inspections, and investigations into foodborne illnesses. Inspections are mostly unannounced, allowing EHS sanitarians to observe food handling practices, check food temperatures, and assess compliance with regulations. Following an inspection, the sanitarian provides the establishment with a detailed report outlining any food safety violations and offering guidance on how to address them effectively.

Challenge we faced by scraping Yelp

Our initial goal was to crawl Yelp to collect a comprehensive dataset of reviews for a representative sample of restaurants across Chicago. By analyzing reviews from a major metropolitan area, we aimed to build a robust model capable of generalizing to a diverse set of establishments. However, our web crawling efforts were stymied when Yelp implemented blocking measures to prevent scraping of their review data **(Figure 2)**. Facing this roadblock, we pivoted our approach and instead targeted Philadelphia as the focus area for our data collection. While a smaller market than Chicago, Philadelphia still provided a sufficiently large and varied restaurant landscape to develop and evaluate our model's performance.

Figure 2: Message that we received from Yelp after being blocked



Sorry, you're not allowed to access this page.

Contact Yelp if you keep experiencing issues.

Yelp Academic Dataset

Yelp, is a social networking platform connecting consumers with local businesses. Featuring a robust review system, users can rate businesses from one to five stars and share detailed experiences. With 32 million users worldwide and 287 million cumulative reviews as of December 31th, 2023, Yelp plays a significant role in the restaurant industry, with 11% of its advertising revenue and 17% of reviews attributed to restaurants. Demographically, Yelp users span various age groups, with 25% aged 18 to 34, 36% aged 35 to 54, and 39% aged 55 and older. In terms of education, 64% hold undergraduate degrees, while 12% possess graduate degrees. Economically, 56% of users earn over $100,000 annually.

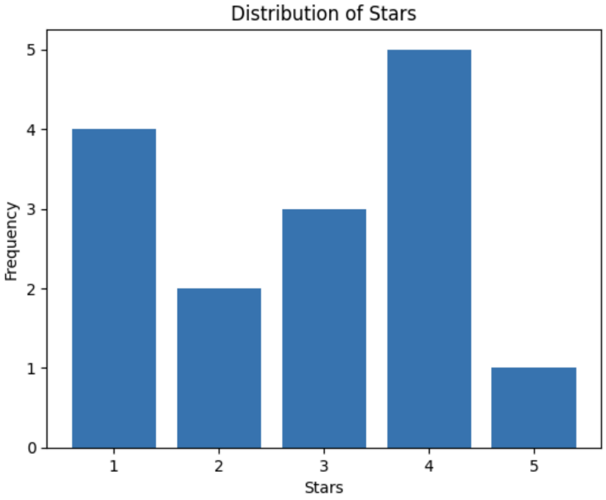
Review distribution shows 53% with five stars, 16% with four stars, 8% with three stars, 6% with two stars, and 18% with one star. At the same time, the distribution of average business ratings is as follows: 16% is rated as 5, 38% as 4, 24% as 3, 16% as 2 and 5% as 1. (Yelp Metrics as of December 31, 2023). **Figure 3** shows the distribution of stars.

The complete dataset of Yelp has different json files. We used three different ones. We included a description, and some characteristics in **Table 1.**

Table 1: Yelp Dataset JSON

Name	Description	Num. of columns	Num. of rows
business.json	Contains business data including location data, attributes, and categories.	14	150,346
review.json	Contains full review text data including the user_id that wrote the review and the business_id the review is written for.	6,990,280	9
user.json	User data including the user's friend mapping and all the metadata associated with the user.	22	1,987,897

Figure 3: Yelp Distribution of stars

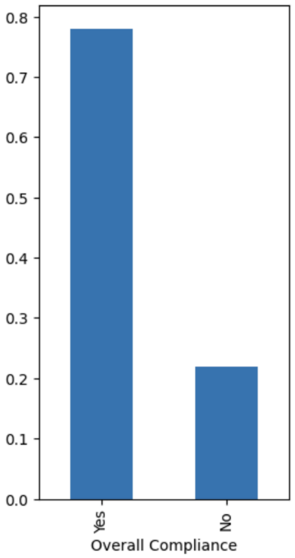


Descriptive Statistics

Merging the datasets

After merging inspections with relevant reviews, our dataset comprised 2,165 inspections across 1,057 unique restaurants, with an average of 7.8 reviews per restaurant, totaling 16,897 reviews. The outcome variable, "Overall Compliance" with health guidelines, indicated a compliance rate of 78%, with 22% non-compliance as we show in **Figure 4.** The dataset was split for training (80%, 1,754), validation (10%, 216), and testing (10%, 202) purposes.

Figure 4: Overall compliance in Pennsylvania



Inspections can happen for a number of reasons. The most common ones are due to protocol (76%), follow-up inspections (15%), operational reasons (5%) and client's actions (4%). In our sample, around 78% of restaurants show compliance towards health inspections. Following **Table 2** there are 10 different types of health inspections in Philadelphia and they follow this distribution regarding the overall compliance.

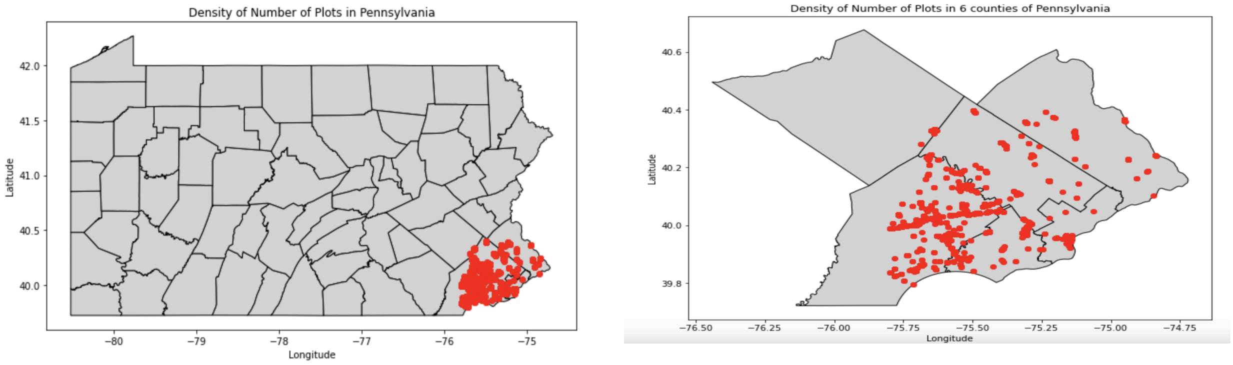
Table 2: Overall compliance and distribution of type of inspection

	Total	***% Yes***	***% No***
All Inspections	***2165***	***78.06***	***21.94***
Regular	1649	75.01	24.98
Follow-up	311	88.42	11.57
Opening	76	96.05	3.947
Complaint	69	82.60	17.39
Change of Owner	40	77.5	22.50
Emergency Response	10	80	20
2nd Follow Up	6	100	0
Type 2 Follow-up	2	100	0
Foodborne Investigation	1	0	100

Health Inspections in Pennsylvania

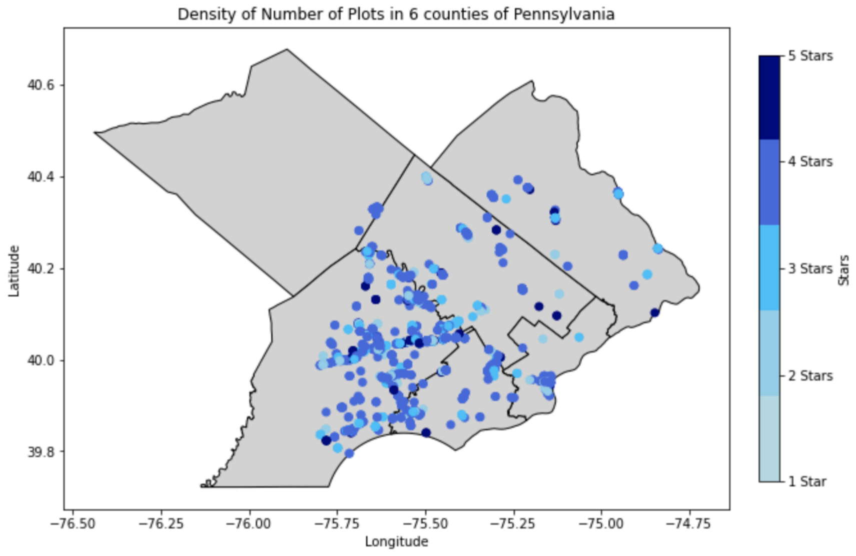
There are 20 jurisdictions in PA in charge of overseeing health inspections. These jurisdictions are heterogeneous in terms of the level of government they belong to: some depend on state-level organizations, others on specific boroughs. Between March 2018 and March 2020, they were in charge of conducting inspections on 1,057 restaurants scattered across 52 different counties in PA. However, as we show in the two elements of **Figure 4,** we only had geolocated inspections for the south east of Philadelphia. These counties' names are: Chester, Delaware, Montgomery, Bucks, Philadelphia and Berks.

Figure 5: Geographical distribution of Philadelphia inspections



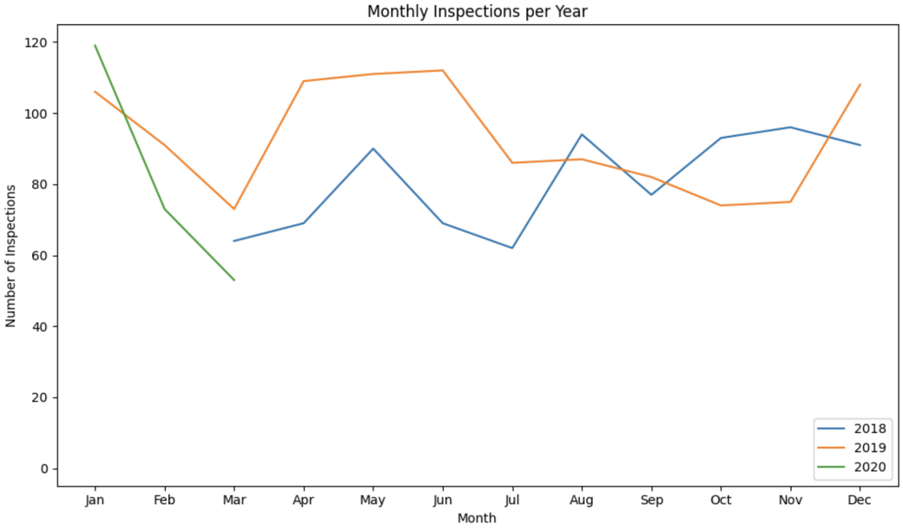
Furthermore, **Figure 6** suggests that there is no discernible geographic pattern in the distribution of star ratings.

Figure 6: Geographical distribution of Philadelphia inspections by star number



We also observe some variation regarding the number of monthly inspections done by all jurisdictions in **Figure 7**. These tend to spike before the holidays (December, January), sharply drop afterwards, and bump up again in the months preceding summer (April through June). Intuitively, this suggests that inspections are conducted before-during periods of high demand.

Figure 7: Distribution of ratings by month (2018, 2019, 2020)



Customer Reviews

People reported impressions of restaurant's products and services can be manifested through numeric ratings (i.e., number of stars) and/or text. We exploit both sources in our dataset. The Yelp dataset also includes a large number of additional features like restaurant categories and attributes, but we found these to be broad and uninformative enough to ignore them further on.

Model Reviews

Data cleaning

For data preprocessing, we performed standard text cleaning techniques such as tokenization, removing newlines and non-alphanumeric characters, and stopword removal. The unit of observation was at the individual review level, with each review mapped to the corresponding restaurant's inspection period. Key feature engineering steps included creating categorical variables for restaurant location across counties in the Philadelphia area, deriving numeric review characteristics like star rating and review count, encoding inspection types, and representing the top 25 cuisine categories as boolean restaurant characteristics. This feature set aimed to capture locational, consumer sentiment, inspection patterns, and restaurant profiles that could inform the likelihood of health code violations.

About the models that we used

We evaluated a wide range of modeling approaches, including traditional machine learning techniques as well as modern neural network architectures. Logistic regression and support vector machine models were trained on the review text data alone as well as with the engineered features. For neural methods, we experimented with recurrent neural networks processing the review text sequences. Additionally, we fine-tuned pre-trained transformer language models like DistilBERT, allowing them to learn from both the raw text and structured data features. This summary can be analyzed in **Table 3.**

Table 3: All the models we used

Neural net approaches	Transformer-based language learning models
● Logistic Regression (text only) ● Logistic Regression (text, features) ● SVM (text only) ● RNN	● distilBERT (text only) ● distilBERT (text, features)

Neural Net Model Architecture

For the neural network model, we began with text preprocessing steps including removing newlines and non-alphanumeric characters from the review data. A TF-IDF vectorizer was used to convert the text into numerical unigram and bigram features, capturing the relative importance of words and word pairs across the review corpus. We experimented with class-weighted random sampling to account for imbalance, but found this did not improve performance.

The model architectures explored included logistic regression models taking just the text features or concatenated with the engineered data features after an initial linear layer. The text-only logistic regression had a linear layer projecting the 2000 TF-IDF features to 2 outputs followed by a sigmoid activation. When including data features, these were concatenated before a final linear output layer. We also evaluated support vector machine models with a linear kernel operating on the text representations. Across all models, a batch size of 16 was used during training. The text preprocessing and modeling iterations aimed to find the optimal architecture and feature inputs for predicting health code violations from restaurant review data.

RNN Architecture

For the recurrent neural network approach, we utilized pre-trained GloVe embeddings of 300 dimensions to represent the review text data. Reviews were truncated or padded to a fixed sequence length of 150 tokens as input to the model. Key hyperparameters included a hidden layer size of 256 units, a batch size of 32, training for 25 epochs with an initial learning rate of 0.001, and a dropout rate of 0.5 to mitigate overfitting. The weight initialization for the hidden layer was set to 0.0.

The model architecture processed each batch by first embedding the input sequences using the GloVe vectors, resulting in a 150 x 300 dimension input. This was passed through an LSTM layer, producing a 256-dim hidden representation. A linear layer then projected this to 2 outputs, which were normalized by a sigmoid activation to generate the predicted probabilities for the binary classification task. This RNN architecture leveraged sequential modeling of the review text to capture semantic patterns indicative of potential health violations.

BERT Model Architecture

For the transformer-based approach, we leveraged the pre-trained BERT encoder model to obtain contextualized embeddings of the review text data. During preprocessing, the first 512 tokens of each review were encoded, with shorter reviews being padded and longer ones truncated. This review embedding matrix of dimensions (batch size x 512) was used as input to the model.

In the text-only architecture, the BERT output embeddings passed through a linear layer projecting to 2 logits, which were then normalized by a sigmoid activation to generate the binary classification probabilities. For the multimodal variant, the BERT text representations were combined with the engineered data features through concatenation and an additional linear layer before the final sigmoid output.

We used a training batch size of 16 and validation/test batch size of 8. No additional sampling was performed. This architecture capitalized on BERT's bidirectional transformer to model the review text, while the multimodal version also integrated the structured data features to inform the health inspection violation predictions.

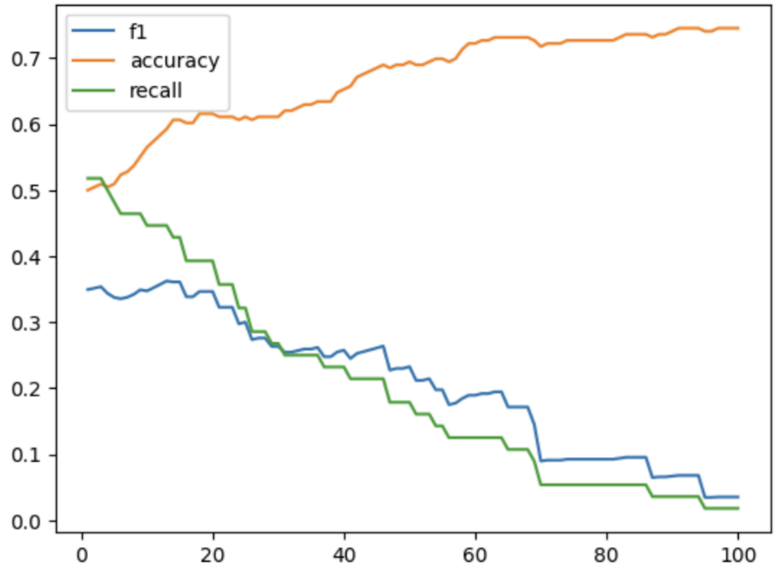
Table 4: Results Table

	By Inspection Period			**By Review**		
Model	Accuracy	F1	Recall	Accuracy	F1	Recall
Logistic Regression (text only)	0.745	0.035	0.018	0.759	0.000	0.000
Logistic Regression (text and features)	0.718	0.032	0.018	0.754	0.016	0.008
Support Vector Machines (text only)	0.741	0.000	0.000	0.759	0.000	0.000
Fine-tuned BERT (text only)	0.806	0.892	1.000	0.786	0.880	1.000
Fine-tuned BERT (text and features)	0.800	0.622	0.539			

Limitations

As we show in **Table 4** and **Figure 8** we ran with some limitations. With a limited number of examples, the model struggled to effectively learn meaningful patterns from the text that could generalize to accurately predict health code violations. Even when incorporating the engineered categorical features like location and cuisine type, the logistic regression approach failed to gain enough predictive power to outperform a naive baseline of predicting the majority class label.

Figure 8: Key Metrics for Logistic Regression (only text)

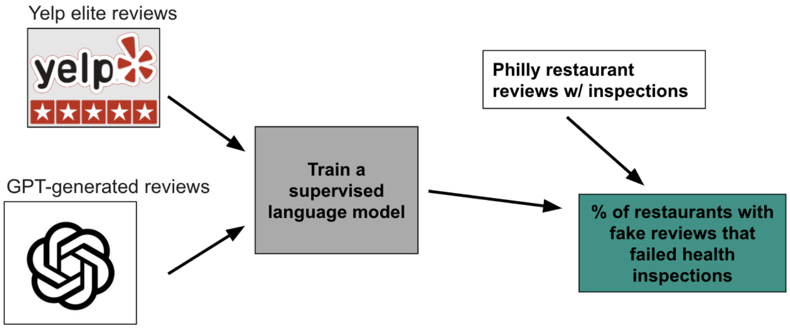


Classifying Fake Reviews

Our new task

Our new task focused on developing a fake review detection model to filter and clean the consumer review data prior to using it for inspection prioritization; it can be seen in **Figure 9**. We constructed a labeled training dataset by treating reviews scraped from GPT as potential fakes or deceptive content, while assuming 'elite' Yelp reviews represented truthful experiences from real consumers. This allowed us to train a binary classification model to discriminate between fake and authentic reviews based on linguistic cues and patterns learned from the data. By deploying this fake review detector, we could remove misleading and unreliable feedback from consideration, ensuring our inspection model learns only from a corpus of genuine crowd sourced signals about restaurant operations and health code compliance (Gambetti and Han).

Figure 9: Our new task


Yelp elite reviews
GPT-generated reviews
Train a supervised language model
Philly restaurant reviews w/ inspections
% of restaurants with fake reviews that failed health inspections

Generating labeled dataset

To create our labeled training data, we took two different strategies for obtaining real and fake review examples. For the real review corpus, we extracted all reviews written by "elite" Yelp members, who have been verified as authentic users by the platform. These elite reviews were labeled as genuine examples of consumer feedback.

On the other hand, we used OpenAI's API to generate synthetic fake reviews by providing zero-shot and "few"-shot prompts to GPT-3.5 and GPT-4.0. The prompts were carefully crafted to instruct the models to produce reviews exhibiting characteristics of fake, deceptive or paid content while maintaining plausibility. The generated text from these large language models served as our labeled fake review data.

To robustly test the performance of our models' ability to distinguish human from computer generated text, we need to extract and use relevant parameters from real reviews. Given the features available to us in the Yelp data, we decided on choosing three variables to be used as inputs for generating reviews: name of restaurant, rating (number of stars), and review length. The restaurant's name and review length are randomly picked from the underlying distribution of the verified data, and for the number of stars we used non-uniform random sampling based on the rating probabilities from the empirical distribution. This process leads to the construction of prompts like the ones shown in **Table 5**.

Table 5: Base prompts using (pseudo) randomly drawn review characteristics

Example #1	**Example #2**
You attended a restaurant named *Lerua's Fine Mexican Food*. You rated your experience with *4* stars out of 5. Write a review of *1248* characters describing your experience.	You attended a restaurant named *Legal Sea Foods*. You rated your experience with *1* stars out of 5. Write a review of *754* characters describing your experience.

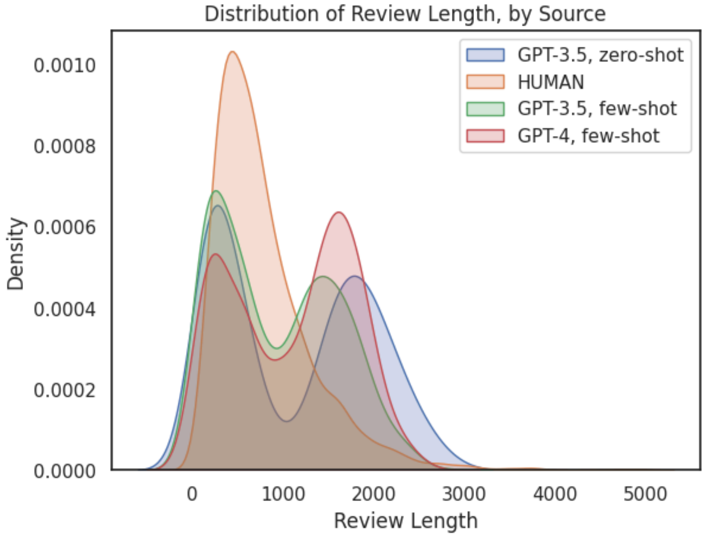
The above are examples of zero-shot prompts, i.e. instructions provided to ChatGPT without including examples. For few-shot prompts, we randomly pick reviews from the verified dataset. Following this process, we arrive at fake reviews of the following nature:

Table 6: Fake reviews from zero/few shot prompting with GPT-3.5/4

Example #1	**Example #2**
Loved my experience at Jimmy J's Cafe! The cozy atmosphere and delicious food made it worth the wait. Highlights were the blueberry brandy with Brie cheese french toast and the unique shrimp & garlic aioli french toast. Can't wait to come back! #Foodie #Deliciousness	Disappointing experience at Baby Blues BBQ Philly. Food was average and service was lacking. Uninspiring flavors and slow service left me unimpressed. Would not recommend. ⭐⭐

At first glance, fake reviews are indistinguishable from real ones in terms of phrasing and level of detail. Some differences that stand out relative to real reviews are the use of hashtags or emojis. However, we are more interested in a general overview of true/fake reviews in terms of shape and substance. To analyze these characteristics for both labels, we overlap the distributions of review lengths (i.e., shape), and ratings (i.e., substance) using one of BERT's sentiment analysis models which translate sentiment in terms of number of stars).

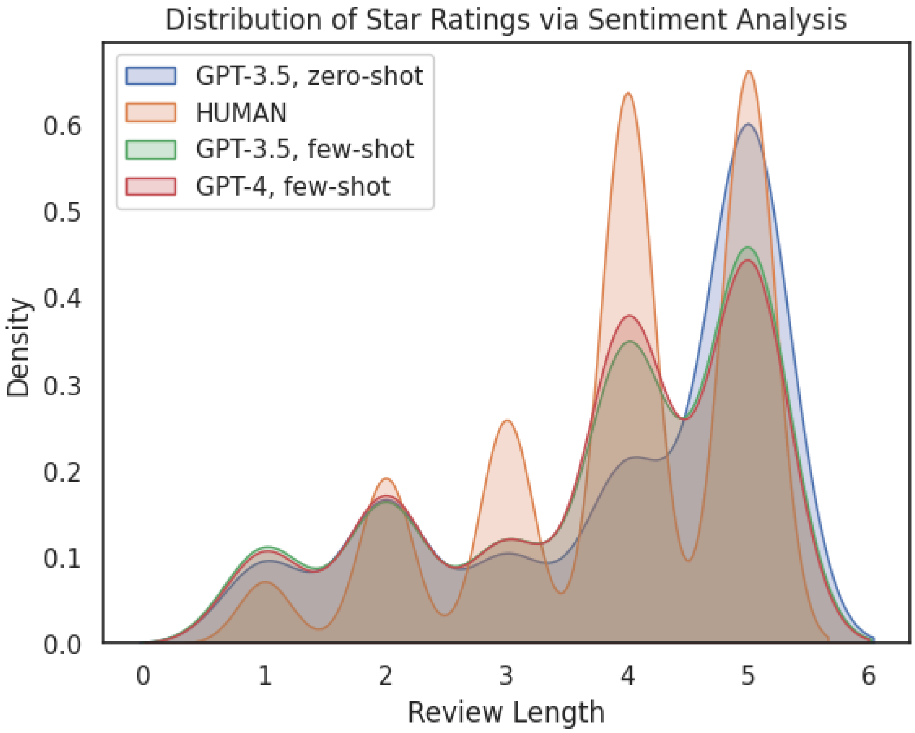
Figure 10: Distribution of Review Length



In terms of review shape, **Figure 10** shows some interesting results. For human-written ones, the distribution is one-modal and it's centered around 900 characters long, a situation that differs for fake reviews. Despite being given clear instructions regarding response length, and accounting for variation in model and prompt type (such as few-shot examples), all GPT-generated reviews are bimodal, with one mode aligning more closely to the mean length of human reviews, while the other one is closer to 2000 characters of length.

Regarding the substance of reviews, we build a number-of-stars rating via sentiment analysis of their content using BERT's *base-multilingual-uncased-sentiment* model. Results are shown in **Figure 11**. There we can see that, relative to GPT-generated reviews, human reviews tend to be much more skewed to positive reviews (i.e., 4-5 stars) than negative ones, possibly due to people taking the time to make a good review when they had a nice experience. A situation that is not reflected in fake reviews, where the proportion of negative reviews is larger.

Figure 11: Distribution of Star Ratings via Sentiment Analysis



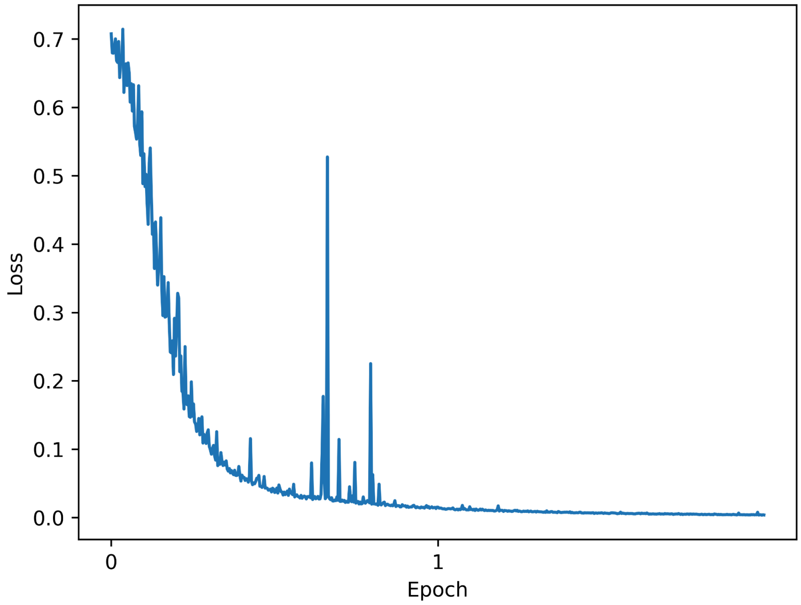
Results

Using these new generated data we test the performance of two models: an RNN and a fine-tuned BERT model. Compared to the results exhibited before, in **Table 7** we see an explosive improvement in terms of F1 and Recall scores, and another significant but smaller one in terms of accuracy. Between both models, the fine-tuned BERT model proves to be better at the task of predicting fake reviews. However, we need to take these results with a grain of salt since we may be experiencing overfitting, as shown by the fast drop in loss in **Figure 9**.

Table 7: Fake Reviews Results

	Per Review		
Model	Accuracy	F1	Recall
RNN	0.792	0.841	0.990
Fine-tuned BERT (text only)	0.998	0.998	0.997

Figure 12: Fine-tuned BERT loss per epoch



Next Steps

After attempting to predict restaurant's health inspection approval using review data and then pivoting to detecting fake reviews, we found considerable room for improvement in this last task. The flexibility that OpenAI's various models have in terms of allowed creativity (temperature) and provided input via fine-tuning prompts for creating fake reviews (e.g., allowing misspelling) allows us to test different generation processes that could lead to more reliable results. Moreover, there may be additional data sources worth exploring, such as demand-sided Google Reviews or supply-sided ones like restaurant online menus, which may provide deeper context about the restaurant for the creation of fake reviews.

Effort Description

Group Member	Assigned Tasks	Effort	Previous/New Skills
Claire Boyd	1. Generated random sample of Chicago restaurants using yelp API 2. Merged/cleaned philadelphia data 3. Handled data loading / pre-processing for pytorch (built classes, decided on architecture) 4. Built Logistic and SVM models 5. Created functions to measure model performance & generalize across model types	1. Medium 2. Low 3. High 4. High 5. High	Model architecture, data cleaning & preprocessing, abstracting processes for replication (creating Classes and functions to use across models)
Raúl Castellanos	1. EDA (inspections/reviews and geographical components) 2. Data stration of verified reviews from Yelp Academic Dataset. 3. First draft of ppt and word doc.	1. Medium 2. High 3. Low	Use of polars for manipulation of big datasets; geopandas.
Jack Gibson	1. Scraper to collect review text from Yelp.com 2. Feature extraction, data cleaning, and pipelining for inspection data and GPT-generated review data 3. Data preprocessing and training/evaluation framework for deep learning models 4. Constructed fine-tuned BERT and RNN-LSTM models 5. Wrote notebooks for deploying/training models using GPU	1. Medium 2. Medium 3. High 4. High 5. High	Model design (PyTorch), fine-tuning transformer models from HuggingFace
Benjamin Leiva	1. EDA (inspections/reviews) 2. Sentiment analysis with BERT model (on original/new data) 3. Generation of fake reviews	1. Medium 2. Medium 3. High	Data Visualization, Interaction with OpenAI's API

Bibliography

- Yelp Metrics as of December 31, 2023 https://www.yelp-press.com/company/fast-facts/default.aspx
- Philadelphia Data https://www.phila.gov/services/permits-violations-licenses/get-a-license/business-licenses/food-businesses/look-up-a-food-safety-inspection-report/
- Food Facility Inspectionsin the City of Philadelphia https://www.phila.gov/media/20190429115924/Food-Facility-Inspections-in-the-City-of-Philadelphia_2019.pdf
- Predicting Restaurant Health Violations Using Yelp Reviews: A Machine Learning Approach https://www.harlanhutton.com/yelp.pdf
- Combat AI With AI: Counteract Machine-Generated Fake Restaurant Reviews on Social Media https://arxiv.org/abs/2302.07731